International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)
ISSN (P): 2249–6831; ISSN (E): 2249–7943
Vol. 12, Issue 1, Jun 2022, 123–134

TRANS STELLAR
•Journal Publications •Research Consultancy

AUTOMATED SYSTEM OF EXPLORATORY DATA ANALYSIS

SRAJAN KUMAR[1], SHASHWAT JAISWAL[2], UTKARSH RASTOGI[3], VEDIC RAGHUVANSHI[4] & SOUMI GHOSH[5]

Department of Computer Science and Engineering, Inderprastha Engineering College

Ghaziabad, U.P., India

ABSTRACT

In today's information-driven business environment, data is the fuel for growth in both data-driven and non-data-driven organizations. Channelized analysis and visualization of data proved to be key factors in improving the business models and the overall performance of any organization hence EDA is one of the key processes which helps organizations to make informed business decisions by examining or understanding the data and extracting meaningful insights from the data by exposing trends patterns and the relationships that are not readily apparent.

By looking at and analyzing the research done previously in the field of EDA, we have identified a few gaps in the process of performing EDA on a dataset. Firstly, the data visualization techniques and ways that were primitively used and no longer contain any significance need to be made more efficient and precise. Another concern is that most of the time and effort while performing EDA goes into data cleaning and pre-processing, after which the actual data analysis and insight gathering begin. To improve the efficiency of the process, this research paper introduces automation in the process of EDA.

KEYWORDS: Automation, Exploratory Data Analysis, Data Visualization, Data Analytics, Data Pre-processing, Data Science

Received: Mar 26, 2022; **Accepted**: Apr 16, 2022; **Published**: May 17, 2022; **Paper Id**: IJCSEITRJUN202215

Original Article

INTRODUCTION

In today's digital world, insights obtained from Exploratory Data Analysis (EDA) are used in strategic business decision-making. EDA is a fundamental procedure that makes use of statistical techniques and graphical representation to obtain insights from data. EDA not only assists with the identification of hidden patterns and correlations among attributes in data but also helps with the formulation and validation of hypotheses from the data. Over the last few decades, interactive visualization strategies have become an integral part of data exploration and analysis techniques.

With a picture being worth a thousand words, academics have proposed several tools and techniques to visualize complex relationships among data attributes using simple diagrams and charts. Whilst some of these visual data analysis tools assist with domain-specific data analysis (for example, analysis of genome sequence data, meteorological data, results of predictive analysis), some other tools focus on general-purpose exploratory browsing of tabular data. In either case, since the beginning of visual interactive data analysis, almost all visual EDA tools perform a few common analytics tasks. In their work, as well as this paper has identified these basic data exploration tasks as sort, filter, aggregate, correlate, group, and derive attributes.

Now, since these repetitive tasks amount to a great deal of time and effort to perform, again and again,

there needs to be some way to increase the productivity and efficiency of the overall process.

OBJECTIVES

- To automate the same repetitive tasks associated with data analysis such as Data cleaning, Data preprocessing, etc.
- To change the traditional way of analyzing the data and using it to make meaningful decisions.
- To reduce the processing period of raw data.
- To promote the reusability of data by uploading it to the cloud so that it can be used as and when required without any preprocessing.

BACKGROUND RESEARCH

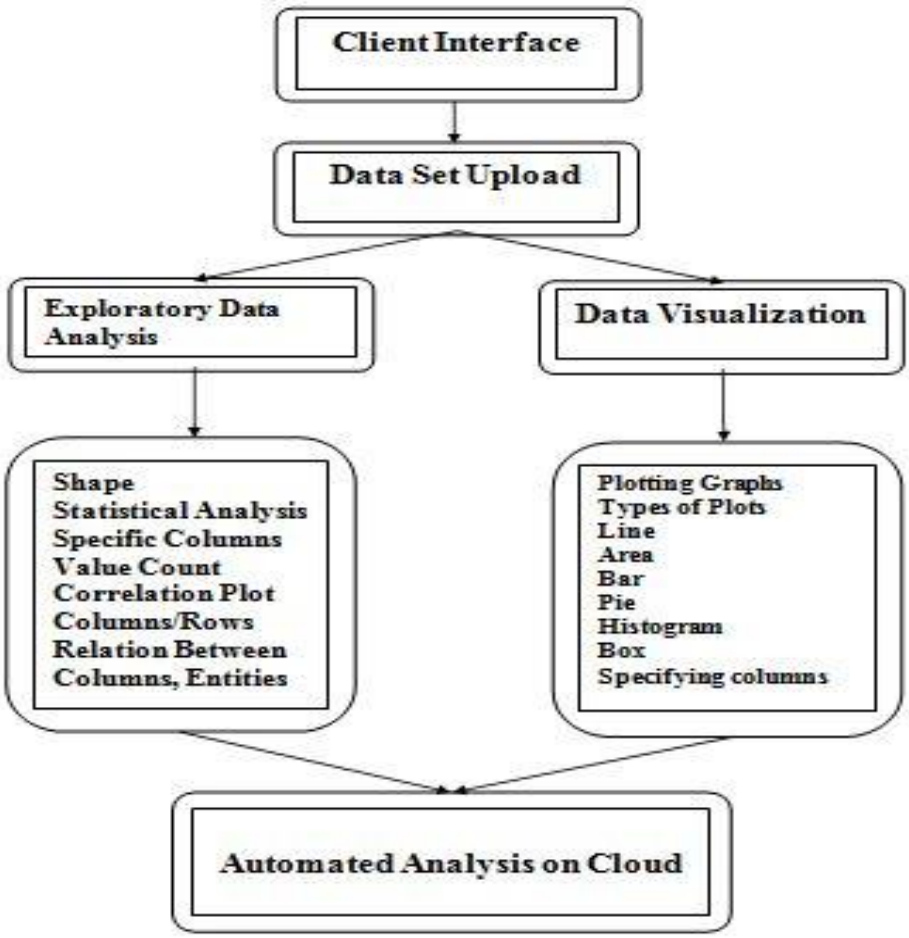
This paper has referenced the following papers ranging from 1984 to 2020 which are summarized as follows:

Cleveland et al [6] suggested various techniques to improve the perception of graphs we generate during EDA which helps in more channelized analysis. Buja et al [3] addressed some of the issues with visualization of high dimensional and large data sets and also proposed various approaches to resolve these issues. Friendly M. et al [8] talked about various types in which data can be displayed using graphs and how insights can be gathered from it. Billard et al [2] emphasized how data can be analyzed and visualized more efficiently so that we may be able to gather that knowledge and information from it which is of practical use and application. Gelman et al [9] highlighted the Bayesian formulation of Exploratory Data Analysis and suggested different approaches to the same. Johnstone I.M. et al [13] addressed some of the statistical challenges which occur while analyzing high dimensional data and suggested some ways to deal with such situations. Chong Ho Yu [5] proposed a conventional conceptual framework of EDA with the help of data mining and resampling with the use of cluster detection, variable selection, and pattern selection. C. Chen [4] talked about the importance of information hidden in the data which can be extracted using meaningful visualization and gathering actionable insights from it. L. Yu et al [14] addressed various issues on time-varying data visualization and proposed some of the methods for automatic animation of such type of data. Lei Yang et al [15] EDA technology gets rapid development, and more and more EDA tools and software come out. The application, function, design flow, and some important development tools of EDA technology are introduced in his paper. S.A. Murphy [18] suggested the use of BI tools for visualization and creating dashboards to support library decision-making. Also proposed were various ways to make this process quick. Idreos S. et al [11] provided a basic overview of various data explorationtechniques which helps in drawing various kinds of conclusions from raw data. J. Wolfe [12] emphasized the importance of data in data visualization as to how to make data interactive and appealing. X. Li et al [20] emphasized the importance of large data visualization and suggested the use of advanced aggregate computation for the analysis of huge data sets. Godfrey P. et al [10] addressed some of the issues with performing EDA onlarge data sets such as the time required to preprocess such large data and many innovative solutions have been proposed. T.J. Brigham [19] introduced the concept of data visualization uniquely and intriguingly and emphasized story-telling about charts and correlations. R.R. Laher [17] informed about ‘Thoth’, a software for data visualization and statistics, discussed various functionalities of this software and how can this be an ideal choice for EDA. Battle L. et al [1] proposed dynamic pre-fetching of data tiles to make the processof EDA faster and as interactive as possible. Also emphasized the importanceof data management in EDA.El Hindi et al [7] discussed VisTrees,

a tool for visualizing and characterizing subgroups in a data set, and also talked about fast indexes for interactive data exploration. Yalcin M.A et al [21] introduced the concepts of expressive tabular data analysis along with the methods to make this process rapid and time-efficient. Rahul Reddy Nadikattu [16] elaborated on the modern techniques of research in the fields of Data Science, Data Analytics, and Data Visualization. Modern techniques included those which are time efficient and consistent. After reviewing these papers, it was found that no research work has been done in the field of EDA to make it a time-efficient process.

METHODOLOGY

System Architecture



MAJOR MODULES AND THEIR FUNCTIONALITIES

Client Interface (Front end): It is the user interface that the user will use tointeract with the system and perform automated data analysis and visualization.

It is designed using streamlit, which is an open-source Python library that makes it easy to create and share beautiful, custom web apps.

Upload Feature: It enables the user to upload a dataset to perform EDA on that data. A user can upload data in

TXT, or CSV format for now but it willbe enhanced to accept data in more kinds of formats.

Core Functionality Modules:

Exploratory Data Analysis: In this module, several features of EDA will be automated for the user just with a single click of a button such as plotting the correlation plot using matplotlib and seaborn, drawing the pie plot by clickingon its button, and carrying out other statistical analysis for more precise and accurate analysis.

Data Visualization: In this module, the various methods and ways for visualizing data are automated for the user like generating various kinds of charts such as bar charts, histogram, area, kde, etc, and that too using any single or multiple attributes at a time. All a user needs to do is to select the type of chart and the attribute as a parameter from the drop-down list.

Database: At the current stage, the user needs to upload a data set every time he/sheneeds to perform EDA so the current progress is lost every time user closes the system but with the help of a database a user will be able to store the data on the database and his/her progress will not be lost. This module will also ensure dataintegrity and security.

Login/Register Page: This module enables a new user to register himself on thesystem and then login into the system to perform EDA in a more secured and customized manner.

Experimental Setup

Pandas is a free Python software library for data analysis and data handling. Pandas provide various high-performance and easy-to-use data structures and operations for manipulating data in the form of numerical tables and time series.

NumPy is a free Python software library for numerical computing on data that can be in the form of large arrays and multi-dimensional matrices. These multi-dimensional matrices are the main objects in NumPy where their dimensionsare called axes and the number of axes is called a rank.

Scikit-learn is a free software library for Machine Learning coding primarily in the Python programming language. It was initially developed as a Google Summer of Code project by David Cournapeau and was originally released in June 2007. Scikit-learn is built on top of other Python libraries like NumPy.

Matplotlib is a data visualization library and 2-D plotting library of Python. You can use Matplotlib to create plots, bar charts, pie charts, histograms, scatterplots, error charts, power spectra, stemplots, and whatever other visualization charts you want.

Seaborn is a Python data visualization library that is based on Matplotlib and closely integrated with the NumPy and pandas data structures. Seaborn has various dataset-oriented plotting functions that operate on data frames and arraysthat have whole datasets within them.

Streamlit is an open source app framework in python language. It helps us createbeautiful web apps for data science and machine learning in a little time. It is compatible with major python libraries such as scikit-learn, Keras, pytorch, latex, numpy, pandas, matplotlib, etc.

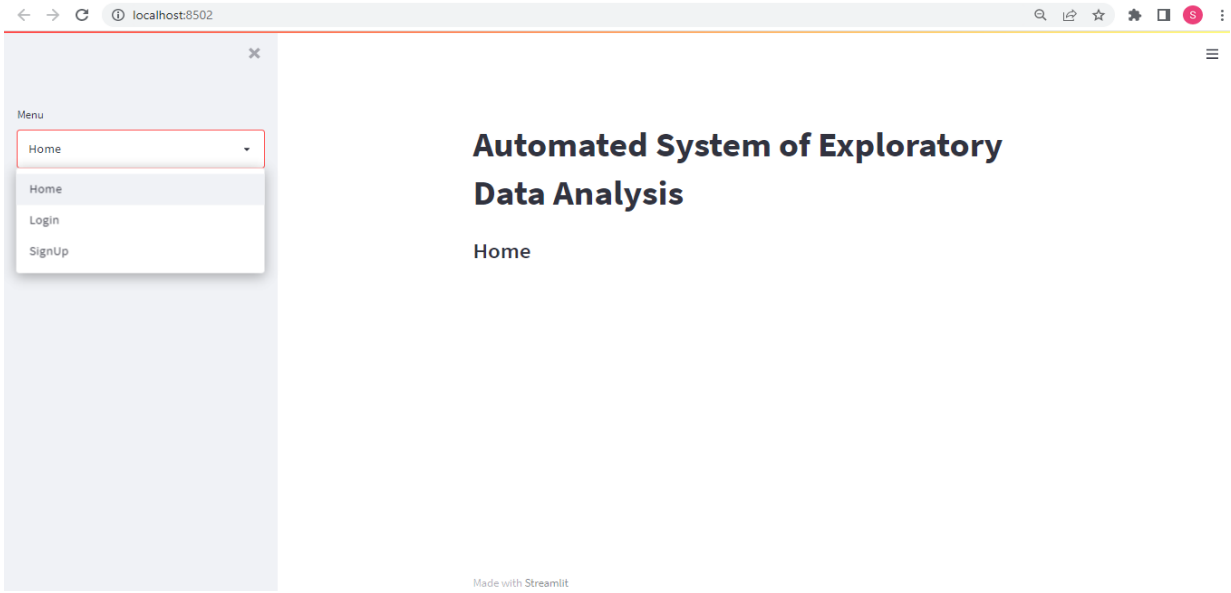
IMPLEMENTATION

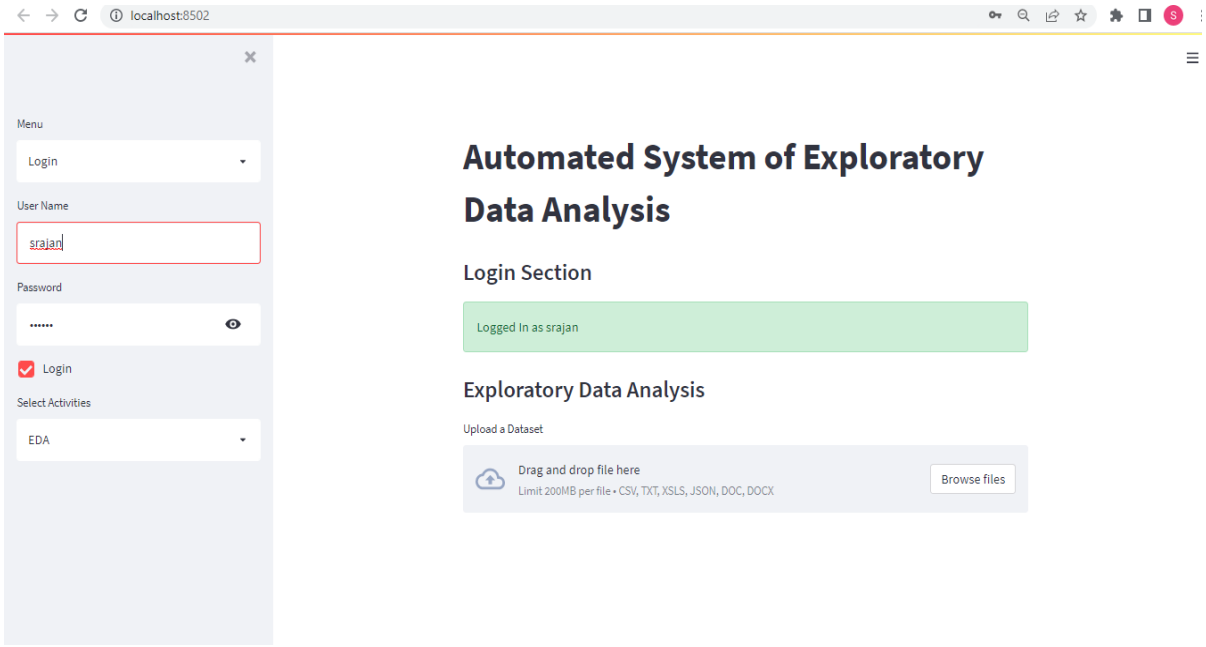
Implementation of Modules/Algorithms

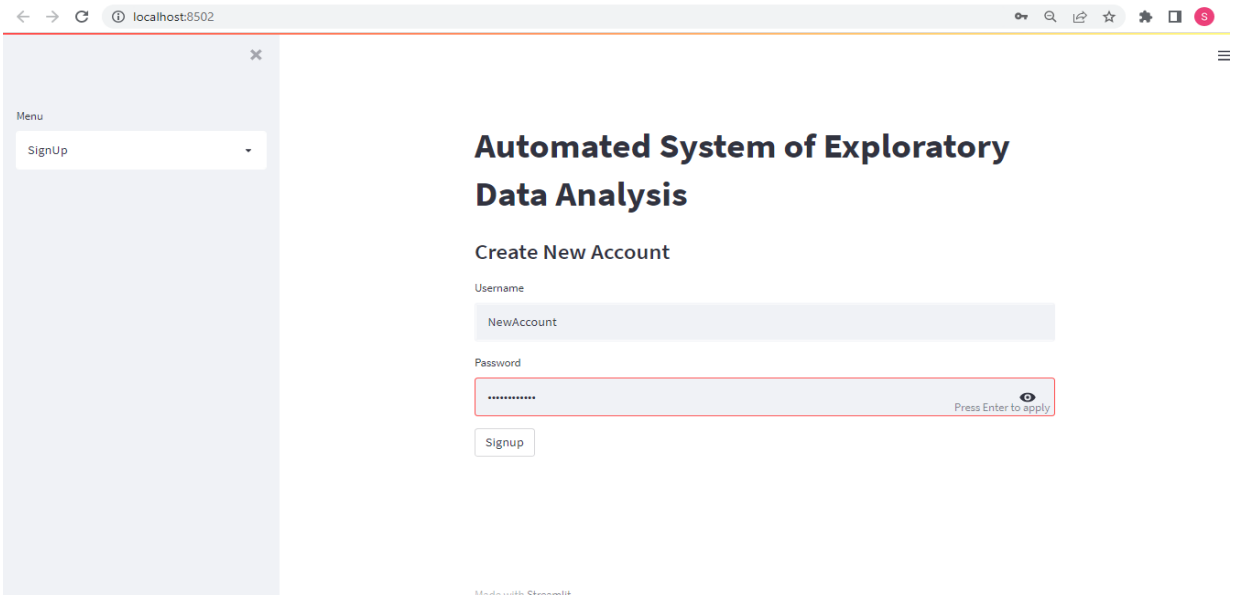
Streamlit for Client Interface

The client interface i.e. the UI designed for the user to interact with the system is designed using Stremlit in python. In other words, the first interaction of the user inthe system will be with the interface designed using streamlit. All the buttons like uploading data, the drop-down for asking the user his/her preference for data analysis, and other UI buttons are designed using streamlit.

For example, the sidebar for selecting the type of activity in the system is designedusing the ‘sidebar’ and ‘selectbox' functions. Similarly, the file upload option wasmade using ‘file_uploader’ function of Streamlit.

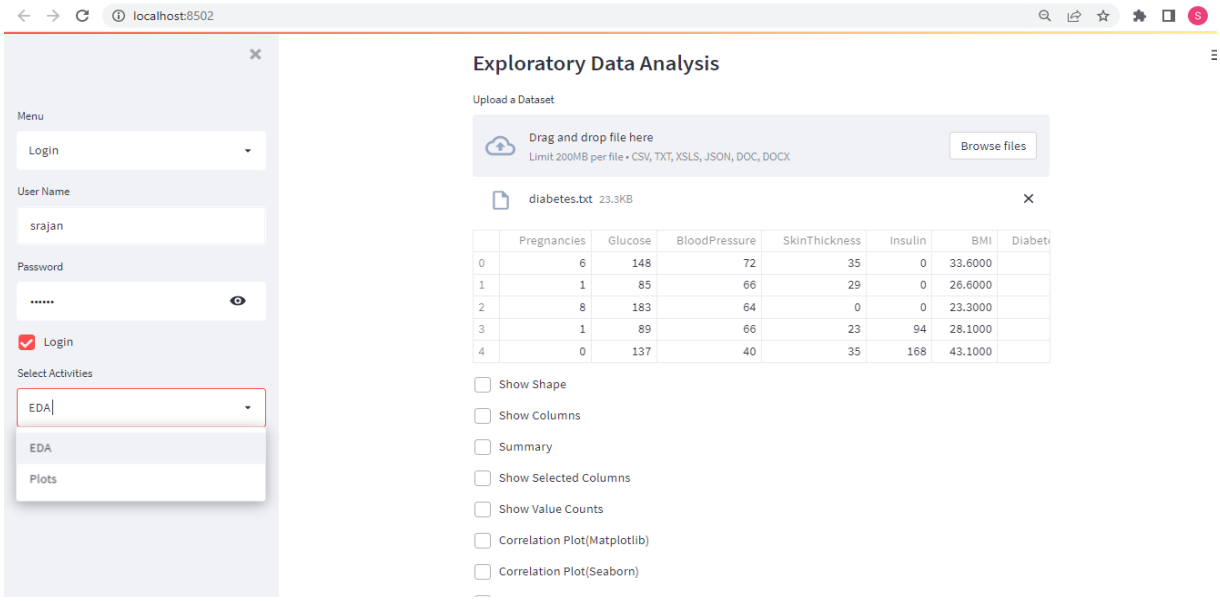






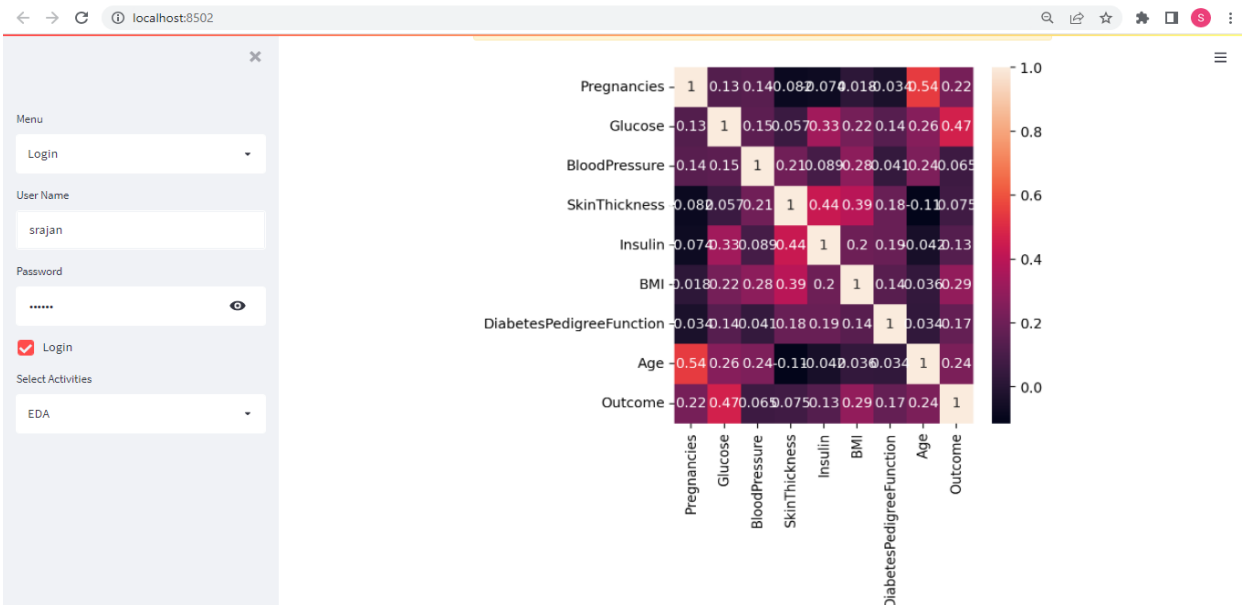
This paper has analyzed the working of the system on a sample diabetes data set. Here are some of the crucial sections of the system with snapshots.

Pandas for Reading the Data. The data which a user uploads using the functionalities made available using streamlit is read by using Pandas. Using this library we can read data inmany formats like txt, CSV, etc.



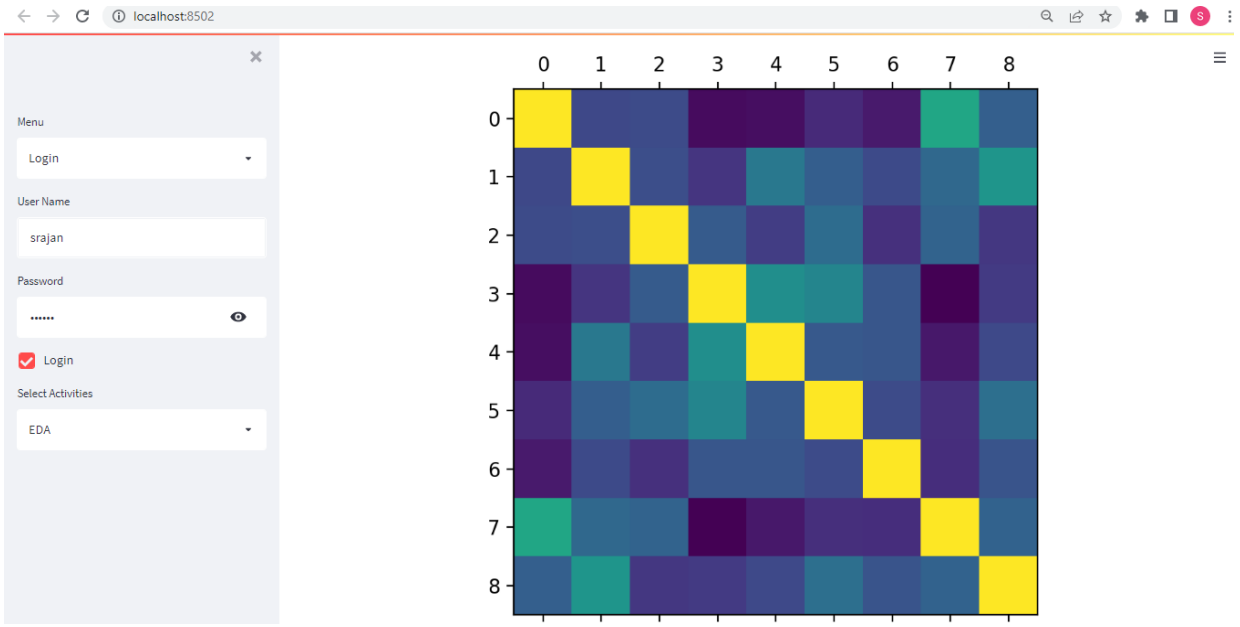
SEABORN FOR GENERATING HEAT MAP

The Heat Map generated as one of the functionalities of our project is generated with the help of the Sea Born library in python. We used Sea Born's function heatmapto generate the heat map for the given data set by specifying the required parameters.



MATPLOTLIB FOR GENERATING VARIOUS PLOTS

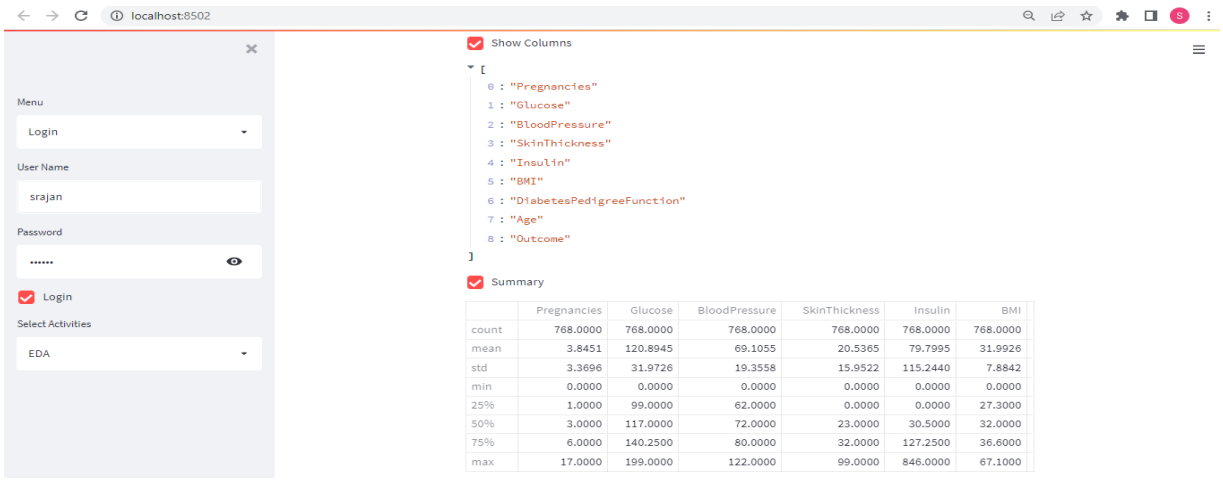
The functions like ‘matshow' are available in the Matplotlib library and were used forgenerating various kinds of plots for the user in the backend. This function was implemented using the ‘pyplot’ class of matplotlib.



NUMPY

Numpy was used in the project to perform numerical calculations in the data and use those calculations to draw various plots and charts in the data visualization section as without these calculations these plots would have been nearly impossibleto draw.

A numpy array provides much more efficient storage and data operations as the arraygrows larger



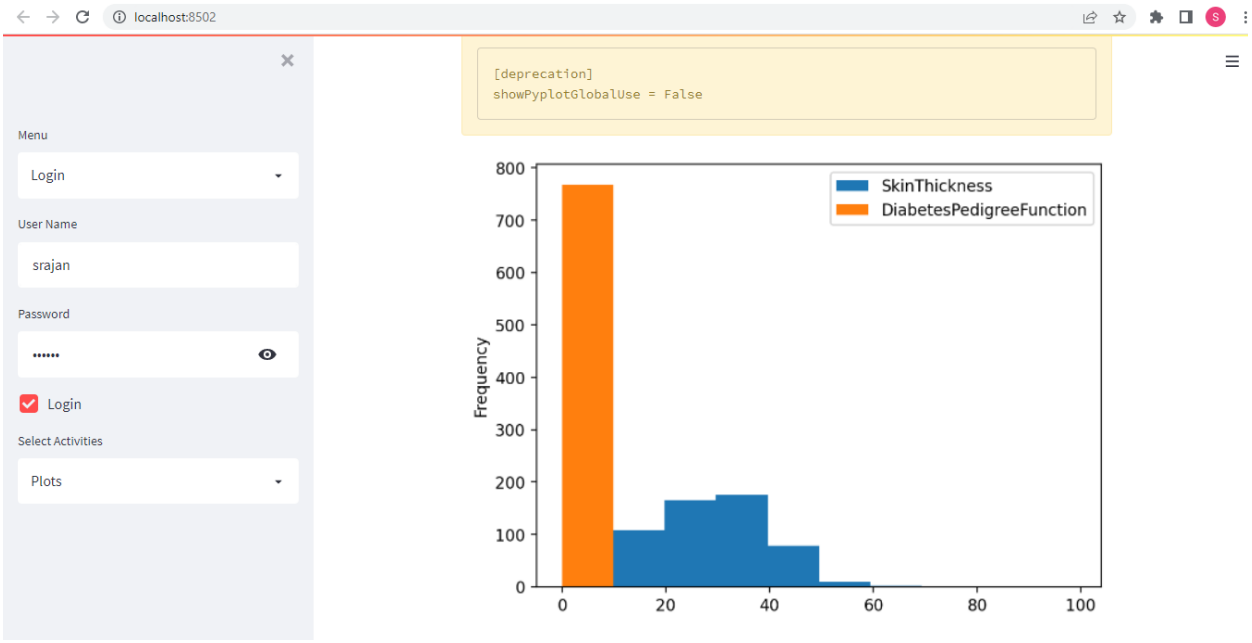
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
count	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000
mean	3.8451	120.8945	69.1055	20.5365	79.7995	31.9926
std	3.3696	31.9726	19.3558	15.9522	115.2440	7.8842
min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
25%	1.0000	99.0000	62.0000	0.0000	0.0000	27.3000
50%	3.0000	117.0000	72.0000	23.0000	30.5000	32.0000
75%	6.0000	140.2500	80.0000	32.0000	127.2500	36.6000
max	17.0000	199.0000	122.0000	99.0000	846.0000	67.1000

DATA VISUALIZATION

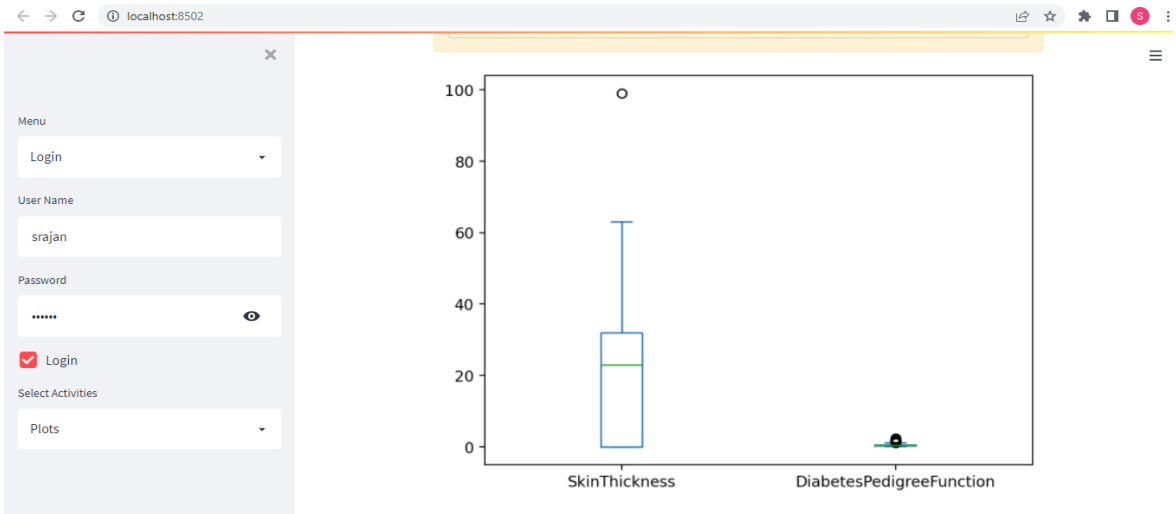
Users can visualize the data in various ways using the drop-down menu feature of the system

GENERATING A BAR PLOT

GENERATING HIST PLOT



GENERATING BOX PLOT



THREATS TO VALIDATE

- For now, this research is suitable for small to medium-sized datasets. The system can be made to accept large datasets in times to come.
- There is a restriction on the types (such as CSV, xlsx, txt) of datasets that can be used with the system, but it can be made compatible with almost all kinds of datasets.
- For now, automation has been applied to a limited number of ways of performing EDA and other data visualization but using a similar kind of approach automation can be introduced for various complex ways as well.

CONCLUSIONS

For this research, various research papers have been studied and analyzed to find that in the field of EDA the traditional methods of analyzing and visualizing data have never been modified and automated to make the process efficient, hence this research has tried to bridge that gap in the process in the form of automating some of the most time-consuming stages in the process of EDA.

For the research, a sample data set has been used to carry out various kinds of analyses on data to authenticate and validate the results obtained using this proposed innovative way of performing EDA.

In the future, the same research can be utilized for additional developments in this proposed system as for now the system works for simple data sets but with more research, the same system can be used to provide multiple features with multi-dimensional data sets as well.

REFERENCES

1. *Battle L. et al., "Dynamic prefetching of data tiles for interactive visualization," Proceedings of the 2016 International Conference on Management of Data, ACM, 1363-1375.*
2. *Billard, L., and Diday, E. - "From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis," Journal of the American Statistical Association, 98, 470–487, 2003.*
3. *Buja, A., Cook, D., and Swayne, D. - "Interactive High-Dimensional Data Visualization," Journal of Computational and Graphical Statistics, 5, 78–99, 1996.*
4. *C. Chen, "Information Visualization," WIREs Computational Statistics, Vol. 2,387-403, 2010.*
5. *Chong Ho Yu- "EDA in the context of data mining and resampling". International Journal of Psychological Research, 3(1), 9-22, 2010.*
6. *Cleveland, W. S., and McGill, R. - "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," Journal of the American Statistical Association, 79, 531–554, 1984.*
7. *El. Hindi et al., "VisTrees: Fast indexes for interactive data exploration," Workshop on Human-in-the-Loop Data Analytics, 5-11, 2016.*
8. *Friendly, M. and Kwan, E.-"Effect Ordering for Data Displays", Computational Statistics and Data Analysis, 43, 509–539, 2002.*
9. *Gelman, A. - "A Bayesian Formulation of Exploratory Data Analysis andGoodness-of-Fit Testing," International Statistical Review, 2003.*
10. *Godfrey P., Gryz J., Lasek P.- "Interactive Visualization of large datasets". IEEE Transactions on Visualization and Computer Graphics, Vol. 28, Issue 8, 2142-2157, 2016.*
11. *Idreos S. et al., "Overview of Data Exploration Techniques," Proceedings of2015 ACM SIGMOD International Conference on Management of data, 277-281, 2015.*
12. *J. Wolfe- "Teaching students to focus on the data in data visualization," Journal of Business and Technical Communication, vol. 29, no. 3, 344-359,2015.*
13. *Johnstone I.M. et al., "Statistical challenges of high dimensional data," Philos.Trans. R. SOC., 367, 4237-4253, 2009.*

14. *L. Yu et al., "Automatic animation for time-varying data visualization," PacificGraphics, Vol. 29, No. 7, 2271-2280, 2010.*
15. *Lie Yang, Li Yao, Xicai Cheng, Bowu Yan- "Research on EDA technology and its related issues". International Conference on Computer Design andApplications, 2010.*
16. *Rahul Reddy Nadikattu, "Research on Data Science, Data Analytics and Big Data", International Journal of Engineering and Science, Vol. 9, Issue 5, 99- 105, 2020.*
17. *R.R. Laher, "Thoth: software for data visualization and statistics," Astronomyand Computing, Vol. 17, 177-185, 2016.*
18. *S.A. Murhy, "Data visualization and rapid analytics: applying tableau desktopto support library decision making," Journal of Web Librarianship, Vol. 7, No.4, 465-476, 2013.*
19. *T.J. Brigham, "Feast for the eyes: an introduction to data visualization," Medical Reference Services Quarterly, Vol. 35, No. 2, 215-223, 2016.*
20. *X. Li et al., "Advanced aggregate computation for large data visualization," Proceedings of IEEE Symposium on large data analysis and visualization, 137-138, 2015.*
21. *Yalcin M.A, Elmqvist N., Bederson B.B.- "Rapid and expressive tabular dataexploration for novices". IEEE Transactions on Visualization and Computer Graphics,24(8),2339-2352,2016.*